A Study of Abbreviations in the UMLS

Hongfang Liu[1], Yves A. Lussier[3], Carol Friedman[2,3]

[1]Computer Science Division, Graduate School and University Center of CUNY
[2]Computer Science Department, Queens College of CUNY
[3]Department of Medical Informatics, Columbia University

Abbreviations are widely used in medicine. The understanding of abbreviations is important for medical language processing and information retrieval systems. The Unified Medical Language System (UMLS) contains a large number of abbreviations. We hypothesized that extracting and studying the UMLS abbreviations can be helpful for understanding the characteristics of abbreviations in medicine. In this paper, we describe a method for extracting abbreviations from the UMLS. We evaluated the method and studied the ambiguous nature of the abbreviations. In addition, the coverage of the UMLS abbreviations in medical reports was studied. Using our method, we extracted 163,666 unique (abbreviation, full form) pairs from the UMLS with a precision of 97.5%, and a recall of 96%. The UMLS abbreviations were highly ambiguous: 33.1% of abbreviations with six characters or less had multiple meanings; the average number of different full forms for all abbreviations with six characters or less was 2.28. The coverage of the UMLS abbreviations in medical reports was over 66%.

INTRODUCTION

In the medical domain, writing favors brevity because time pressures often prevent medical specialists from describing clinical findings fully[1;2]. Many medical words and phrases are long, and abbreviations are a convenient way to shorten them. Abbreviations can take several forms[1;3]:

- Truncating the end, e.g. *adm* for *administration* (or *administrator*),
- First letter initialization, e.g. *AAA* for *abdominal aortic aneurysm,*
- Opening letter initialization, e.g. *Al* for *aluminum,*
- Syllabic initialization, e.g. *BZD* for *benzodiazepine,*
- Combination initialization, e.g. *ad lib* for *ad libitum,* and
- Substitution initialization e.g. *ASD I* for *Primum atrial septal defect* ; *Fe* for *iron.*

According to Bloom[1], abbreviations, which are derived from different types of initialization and which also can be pronounced, are called acronyms. In this paper, we do not distinguish between acronyms and other types of abbreviations.

The abbreviation problem has been shown to affect knowledge-based systems such as natural language processing (NLP) systems and information retrieval (IR) systems in medicine[4-6]. To understand the underlying meaning of abbreviations in a specific domain requires knowledge of that domain. But to manually build the domain knowledge requires a large amount of human effort. There are several reasons. First, the growth of the size of abbreviations is rapid, and it is time consuming to keep knowledge-based systems up to date. Cheung[7] shows that the acronyms used in clinical trials of cardiology alone increased from 200 in 1992 to 2,300 in 1998. Secondly, many abbreviations are ambiguous: one abbreviation can have several full forms (e.g. *Ca* for *cancer* or for *calcium*); and some abbreviations have the same spelling as general English words (e.g. *TOP* for *termination of pregnancy*). In order to have an accurate NLP or IR system, a comprehensive abbreviation knowledge base needs to be built and updated periodically, and a method to resolve ambiguous abbreviations is also needed.

Yoshida and colleagues[3] built a workbench for the construction of a protein abbreviation dictionary. Yu and colleagues[8] developed a method to identify the full form of an abbreviation from parenthetical expressions in scientific articles. The method to extract abbreviations from parenthetical expressions described in this paper is based on their work. In this paper, we used the Unified Medical Language System (UMLS)[9] which was developed by the National Library of Medicine (NLM) to systematically obtain an abbreviation knowledge base. We studied the ambiguous nature of the UMLS abbreviations and the coverage of the UMLS abbreviations in medical reports.

BACKGROUND

The goal of the UMLS is the integration of various vocabularies pertaining to biomedicine. The META (Metathesaurus)[9] is one component of the UMLS. It contains information about biomedical concepts and terms from many controlled vocabularies. The META is organized by concept or meaning, and is

1067—5027/01/$5.00 © 2001 AMIA, Inc.

produced by automated processing of machine-readable versions of the source vocabularies, followed by human review and editing by subject experts. Each concept in the META has a unique concept identifier which itself has no intrinsic meaning, and each unique concept name in the META has a unique string identifier. A concept name and all its variants (which differ from the concept name in upper-lower case and minor variations) are grouped together as a term and a preferred name is chosen for each term. Different concept names with the same meaning are linked to the same concept identifier. In our study we only considered English language concept names. The UMLS Specialist lexicon, an English language lexicon, is another component of the UMLS. The Specialist abbreviation list contains 10,410 unique (abbreviation, full form) pairs in the 2000 version of the UMLS. Some of them are general English abbreviations, for instance, *anal* for *analysis*.

In the META, the names that contain abbreviations are treated as synonyms of the names that contain their full forms, and therefore they are assigned the same concept identifier. For instance, *ERV* and its full form *expiratory reserve volume* are both listed as one of the names of the same concept. Some concept names actually include the abbreviation together with the full form, e.g. *expiratory reserve volume (ERV)* and *ERV - expiratory reserve volume*. *ERV* by itself is also listed as an abbreviation in the Specialist abbreviation list. However, not all abbreviations in the META have a corresponding entry in the Specialist abbreviation list and visa versa. For instance, *TTTS*, which stands for *twin to twin transfusing syndrome* in the META has no entry in the Specialist abbreviation list while *APT*, which stands for *aminopropylisothiuronium*, is in the Specialist lexicon abbreviation list but not in the META.

The New York Presbyterian Hospital (NYPH) Clinical Data Repository[10] provides a central pool into which computer applications can store their data, and from which they can retrieve information placed there by other applications. The repository contains data, which can be narrative data as well as coded. The narrative data contains reports, such as discharge summaries, radiology reports, pathology reports, etc. In these reports, abbreviations are typically capitalized, whereas general English words are typically written in mixed-case or lower-case. In addition, some abbreviations and their full forms are both included using parenthetical expressions. For instance, in the sentence *Preoperative evaluation also included non-invasive flow studies (NIFS),* NIFS is used to represent *non-invasive flow studies*.

METHODS

In this section, we describe the program that extracts abbreviations from the UMLS, the method used to evaluate the program, the method used to study the abbreviation ambiguity problem, and the method used to study the coverage of the UMLS abbreviations. The 2000 version of the UMLS was used.

The extraction program was developed based on manual observation of a training-set. The training-set contained 36,899 concept names, which were concept names in English whose concept identifiers contained the prefix C000. The output generated by the extraction program was a list of (*abbreviation, full form*) pairs. The program was developed to handle the following three cases.

Case 1: An abbreviation and the phrase containing its full form are connected by a dash.
In this case, the abbreviation appears on the left side of the dash and the phrase on the right side. The full form can be the whole phrase, e.g. *AV - aortic valve* or a sub-string of the phrase, e.g. *AV – arteriovenous fistula* or *AV - abnormal atrioventricular connection* (the full form of an abbreviation is underlined).

Case 2: An abbreviation and its full form are included in a parenthetical expression.
In this case, the abbreviation appears inside the parentheses or immediately to the right. In the former case, the full form is a rightmost sub-string of the phrase to the right of the parentheses, e.g. *insertion of intrauterine device (IUD)*. In the latter case, the full form is a whole phrase included inside the parentheses, e.g. *CAD (coronary artery disease)*.

Case 3: An abbreviation and its full form occur in different concept names associated with the same concept identifier.
There are two types of abbreviations defined in this case. The primary type occurs when the abbreviation and its full form occur as two different concept names associated with the same concept, e.g. *ADP* and *adenosine diphosphate*. The derived type is derived from the primary type. For instance, we derive two abbreviation pairs *(abd, abdominal)* and *(cav, cavity)* from a primary type abbreviation pair *(approach through abd cav, approach through abdominal cavity)*.

Initially, the META was processed using the extraction program, and a list of pairs was generated that was subsequently merged with the Specialist abbreviation list to obtain a preliminary UMLS abbreviation list. The preliminary list was processed to remove subsumed pairs: a pair (*abbreviation, fullform2*) is a subsumed pair of (*abbreviation, fullform1*) if each word in fullform2 can be matched to an equivalent portion (either an equivalent word or a full form of that word) in *fullform1*. Two words are considered to be equivalent if there are same or have the same base form in the Specialist lexicon. For instance, in the following, (b) and (c) are two subsumed pairs of (a): *ischaem* is an abbreviation of *ischaemia*, and *ischemia* and *ischaemia* have the same base form in the Specialist lexicon.

(a). (*AMI*, acute mesenteric ischaemia)
(b). (*AMI*, acute mesenteric ischemia)
(c). (*AMI*, acute mesenteric ischaem)

In the abbreviation list, some abbreviations can have several variant forms. For instance, the abbreviation of *American Medical Association* can be written as *AMA* or *A.M.A.*; the abbreviation of *Coenzyme A* can be *CoA* or *coA*. In the following studies, we did not distinguish between abbreviations if they differed in case or punctuation only.

The evaluation studies described in this paper consists of three parts: one part evaluates the method that finds abbreviations; the second part studies ambiguous abbreviations in relation to the number of characters; and the third studies abbreviation coverage of the UMLS in relation to the number of occurrences in the reports.

In order to evaluate the extraction program, we obtained a test set consisting of all UMLS preferred names for 200 randomly selected concept identifiers that were not in the training-set. We considered preferred names because the other names were only variants of the preferred names. A human expert was asked to manually determine abbreviations and denote all (*abbreviation, full form*) pairs. We ran the extraction program using the test set to automatically derive all (*abbreviation, full form*) pairs. To reduce human error, the gold standard was derived by having the same expert re-derive the list of pairs using both the pairs determined by the expert and the pairs determined by the extraction program[11].

For the second part of our evaluation, we hypothesized that ambiguity associated with abbreviations was related to the number of characters in the abbreviations, and that abbreviations with more characters tended to have fewer different full forms than those with fewer characters. We conducted an ambiguity study on a subset of the UMLS abbreviation list: we removed all punctuations from each abbreviation, and if the resulting abbreviation had less than 7 characters, it was included in the subset. In our study, if an abbreviation had multiple full forms, it was considered ambiguous. We computed the average number of full forms in the subset. For abbreviations with the same number of characters, we computed the percentage that were ambiguous, the average number of full forms, and the variance.

For the third part of the evaluation, we studied the coverage of the UMLS abbreviations in medical reports. We used a test set of reports to generate two abbreviation sets, A and B. The test set consisted of reports of patients admitted during 1998 at NYPH in the following domains: discharge summary, radiology, neurophysiology, pathology, GI endoscopy, Ob/Gyn, cardiology, and surgery. The set A was obtained from the test set by extracting (*abbreviation, domain, full form*) tuples, where the *abbreviation* and its *full form* were defined in a parenthetical sentence from reports in the *domain*, with the restriction that the *abbreviation* consisted of 2 to 6 characters. The set B was obtained by using a program to extract a collection of upper-case words ranging from 2 to 6 characters from mixed-case sentences in the test set. We then obtained a preliminary set of B by selecting 40 words randomly from the collection for each domain, with the restriction that no word appeared in multiple domains (to avoid multiple occurrences of a popular abbreviation in the coverage study). For each word in the preliminary set of B, we randomly selected a mixed-case sentence from a report in the corresponding domain that contained that word. All (*word*, *domain*, *sentence*) tuples were presented to a human expert. For each tuple, the human expert used all possible sources (the expert's knowledge, abbreviation dictionaries, WEB sites, etc) to determine if *word* occurred in the sentence as an abbreviation; if it did, the expert supplied the corresponding *full form*, and the tuple *(word, domain, full form)* became an entry in the abbreviation set B.

For each of the abbreviation sets A and B, we first attempted to automatically map the abbreviation and its full form to the UMLS abbreviation list. For those that could not be mapped automatically (because of typos in the supplied full forms or different word orders etc), we manually searched for them in the UMLS abbreviation list. We computed the ratio of the number of matches against the total number of abbreviations.

We hypothesized that the frequency of abbreviations in the reports was related to the UMLS abbreviation list coverage. We computed ratios associated with five different frequency ranges: I (less than 5), II (between 5 and 10), III (between 10 and 20), IV (between 20 and 50), and V (over or equal to 50). The frequency of an abbreviation is the number of occurrences of that abbreviation in the test set. The ratio for each range consisted of abbreviations that had the correct full forms in the UMLS divided by the total number in that range.

RESULTS

There were 137,850 UMLS concept names containing abbreviations that were obtained using the extraction program. The number of unique (*abbreviation, full form*) pairs was 154,444. Among them, 6,567 pairs were obtained from concept names containing dashes, and 1,097 pairs from names containing parenthetical expressions. The number of primary abbreviations (i.e. abbreviations that were original concept names) was 117,149; the number of derived abbreviations (i.e. abbreviations found within concept names) was 30,714. Note some pairs were counted multiple times because they were derived differently. For instance, the pair (*AMP*, adenosine monophosphate) was derived two ways: a) from two concept names (*AMP* and adenosince monophosphate) and b) from a concept name containing a dash (*AMP* - ddenosine monophosphate). After combining this with the Specialist abbreviation list and removing subsumed pairs, we obtained 163,666 unique pairs.

The test set contained 617 preferred names. Among them, 121 concept names were abbreviations as determined by the gold standard. Based on the gold standard, the recall and the precision of the program was 96% and 97.5%, respectively.

In the ambiguity study, there were 16,855 abbreviations in the set. We found that 33.1% of them had multiple full forms. The average number of full forms for abbreviations with less than 7 characters was 2.28. Table I lists the results with respect to the number of characters: *Len* is the number of characters in the abbreviation; *Num* is the number of abbreviations; the number in parentheses is the percentage of ambiguous abbreviations; *Avg* is the average number of full forms; *V* is the variance of the number of full forms.

In the coverage study, there were 364 tuples in the set A; 241 (66.2%) were mapped to the UMLS abbreviation list. The abbreviation set B contained 270 tuples (84.4% of the preliminary set of B); 185 (68.5%) were mapped to the UMLS abbreviation list. Table II lists the results for the sets A and B with respect to the domain: *Num* represents the number of reports; the number in parentheses is the percentage of abbreviations that have matches in the UMLS abbreviation list. Figure 1 lists the results for the sets A and B with respect to the five frequency ranges. Only 30% of the ones occurring less than 5 times were found in the UMLS whereas 80% were found for those occurring more than 50 times.

Len	*Num (%)*	*Avg*	*V*
1	26 (100)	52.6	6.48
2	596(81)	10.9	0.59
3	4137(54)	3.05	0.06
4	5051(27)	1.64	0.03
5	3777(21)	1.41	0.01
6	3268(20)	1.33	0.02

Table I: The ambiguity study results with respect to the number of letters in the abbreviations

Domain	*Num*	*A (%)*	*B (%)*
Neurology	2,758	13(46)	40(70)
Pathology	102,933	132(64)	33(70)
Discharge	23,651	86(72)	33(55)
Ob/Gyn	12,198	3(0)	29(59)
Radiology	306,587	41(78)	33(82)
GI Endoscopy	6,121	3(67)	40(75)
Cardiology	123,799	21(67)	37(76)
Surgery	39,333	65(62)	25(56)

Table II: The coverage study results with respect to the domain

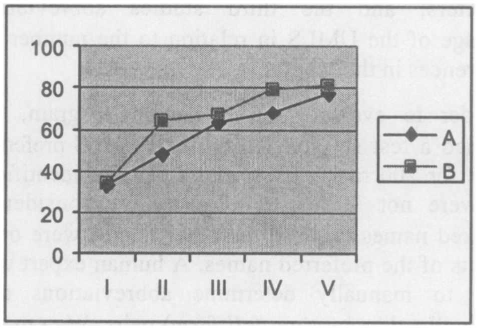

Figure 1: The UMLS abbreviations coverage result for the sets A and B with respect to frequency; the Xaxis represents the range and the Y-axis represents the ratio.

DISCUSSION

We found that most abbreviation pairs (around 90%) in the UMLS abbreviation list were extracted by matching different names of the same concept. Because we found many abbreviations by utilizing the UMLS, we benefited from the expert knowledge that was required to group the different concept names and their abbreviations during the building and updating of the UMLS.

The extraction program performed well in the current evaluation. Most of the false positive errors were caused by the ambiguous nature of the English words, for instance, *removal of eye* and *removal of eyeball* are the same concept in the META, where the pair (*removal of eye, removal of eyeball*) was considered as a primary type of abbreviation and (*eye, eyeball*) as a derived type. Some of the false negative errors were abbreviations containing substitution initialization e.g. (*FX, fracture*).

We found that many abbreviations in the UMLS were very ambiguous, especially those with a small number of characters. For example, there were over 80 different full forms (disregarding the textual variants) for the abbreviation *PA*. The ambiguity of an abbreviation depended on the number of letters it contained: ones with fewer characters were more ambiguous.

In narrative medical reports, the abbreviation ambiguity problem may be more complicated than this study determined. There are different cases of ambiguities concerning biomedical abbreviations. The case where an abbreviation has multiple medical full forms (e.g. *AML* for *acute myeloblastic leukemia* or *angiomyolipoma)* was measured in the current evaluation study. A different case that occurs when the spelling of a medical abbreviation is the same as that of a general English word (e.g. *ALL* for *acute lymphoblastic leukemia*), and another case that occurs when a medical abbreviation is the same as that of a general English abbreviation (e.g. *ASAP* for *atypical small acinar proliferation* or for *as soon as possible*), were not evaluated in this study. In order to have an accurate system, a systematic method to handle ambiguous abbreviations is needed. Some abbreviations can be disambiguated from general English words when abbreviations appear capitalized in mixed-case sentences from reports. In our study, we found that 84.4% of words in the preliminary set of B were abbreviations. Some ambiguous abbreviations can be disambiguated according to domains, but many cannot be. For example, in the set A, *LCA* had three full forms in the pathology domain: *leukocyte common antigen, left coronary artery,* and *lymphocyte common antigen*.

We found the coverage of the UMLS abbreviations was related to their frequency of occurrence in the reports: ones with higher frequency were more likely to be in the UMLS abbreviation list. We also hypothesized that the coverage of the UMLS abbreviations is related to the domain; there was a higher ratio of abbreviations in radiology than in the other domains. Since the size of the abbreviation sets A and B was small and unbalanced for each domain, we could not prove this hypothesis. In order to prove it, a large abbreviation set with a balanced distribution in each domain would be needed.

There are some limitations to this study. First, the extraction program only extracts those UMLS abbreviations whose full forms can be found in names with the same concept identifier. In the evaluation study of the extraction program, we asked the expert to include only those cases. Secondly, we used the outcome of one human expert to derive the gold standard in the evaluation study and the coverage study. It was very time-consuming to derive the abbreviation set B because the expert had to spend more than 15 hours to annotate 320 sentences. Also, in the ambiguity study, ambiguity was determined by the number of different full forms, and not by different meanings. If two full forms of an abbreviation were synonyms, they were counted twice.

Future work will be to: compare the coverage of the UMLS abbreviation list with other abbreviation lists, find a systematic way to improve the manual annotation speed, and develop a method to disambiguate ambiguous abbreviations .

CONCLUSIONS

The UMLS contains numerous abbreviations. We developed a method that systematically extracted those abbreviations, and the method performed well, with a recall of 96% and precision of 97.5%. We found that the UMLS abbreviations were highly ambiguous, particularly those with fewer characters. More work is needed to disambiguate them; this would improve the accuracy of NLP and IR systems. In addition, we found that UMLS has a good coverage of abbreviations in medical reports.

Acknowledgment This study was supported in part by grants LM06274 from the NLM and ILS-9817434 from the NSF.

Reference List

1. Bloom D: Acronyms, abbreviations and initialisms. *BJU Internationa*l 2000:86:1-6.
2. Luxton T, Al-Qassab H: Better use of abbreviations - a lessom from a stroke unit. *Medl Edu* 2000;34: 965-965.
3. Yoshida M, Fukuda K, Takagi T: PNAD-CSS: a workbench for construction a protein name abbreviation dictionary. *Bioinformatics* 2000;16: 169-175.
4.Friedman, C. A Broad Coverage Natural Language Processing System. Proc AMIA Symp 2000: 270-274.
5. Federiuk C: The effect of abbreviations on MEDLINE searching. *Acad Emerg Med* 1999;6: 292-296.
6. Nadkarni P, Chen R, randt C: UMLS concept Indexing for Production Databases: A Feasibility Study. *JAMIA* 2001;8:80-91.
7. Cheung T: Acronyms in clinical traials in cardiology -- 1998. *Am Heart J* 1999;134: 726-765.
8.Yu H, Hripcsak G, Friedman C: Mapping abbreviations to full forms in electronic articles. *JAMIA* 2001;Submittal.
9.UMLS. US Dept of Health and Human Services, NIH, NLM. 2000 Edition.
10. http://www.cpmc.columbia.edu/resources.
11.Friedman CP, Wyatt J: *Evaluation Methods in Medical Informatics*, New York, Springer; 1997.